基于微博大数据的高考舆情研究：热度与情感成分的时空变化及宏观因素影响分析

董锦程 [1,2] 艾钰恒 [3,4] 朱廷劭 [1,2]

[1] (中国科学院心理研究所 北京 100101)

[2] (中国科学院大学心理学系 北京 100049)

[3] (中国科学院自动化研究所 北京 100190)

[4] (中国科学院大学人工智能学院 北京 100049)

摘要 本研究基于微博大数据文本内容，分析了 2019 年至 2024 年间中国大陆各省级行政区有关高考的舆情热度和情感成分变化，并探讨了这些变化与宏观经济指标的关系。研究采用文本条目数量作为热度指标，结合百度指数进行分析，结果表明，高考相关舆情热度在不同年份和省份间存在差异，且与地方经济指标如人均地区生产总值和地方财政教育支出密切相关。而借助林萃分析系统和大连理工情感词典的情感成分分析显示，高考相关文本在不同时间段和地区表现出不同的成分特征，并与一些现实因素相关联。本研究为理解高考舆情的热度与情感成分变化规律，及其与社会经济因素的关系，提供了新的大数据视角。

关键词 大数据；社交媒体；微博；高考；热度数据分析；情感成分分析

引言

近年来，大数据技术的快速发展为社会科学研究提供了革命性工具，尤其是在社交媒体数据的分析领域，极大地扩展了研究范围与深度(杜治娟等, 2017; 吴胜涛等, 2023)。例如，朱廷劭等(2013)利用社会媒体数据提出了群体事件风险预警模型，展示了大数据在动态社会监测中的实际应用。黄峰等(黄峰等, 2020)则通过研究经济发展与集体情绪、道德的关系，证明了大数据在社会经济现象分析中的潜力。

微博(weibo.com)作为中国大陆最具影响力的社交媒体平台之一，因其广泛的用户基础和迅速的信息传播能力，已成为获取公众情绪和行为数据的关键来源，为国内学界广泛采用(常建霞、李君轶, 2021; 陈兴蜀等, 2020; 代一方, 2023; 范晓磊, 2023; 肖嘉锐等, 2021)。在国际学界，基于微博数据的研究同样值得关注：Zhao 和 Wang(2023)通过分析疫情期间的微博数据，揭示了公众与政府媒体之间的框架互动；Xiao 等(2018)的研究通过微博数据，初步实现了对影响市民生活的城市内涝灾害的实时监控。

以微博为代表的社交媒体已成为公众表达独立话语的重要平台。常建霞和李君轶(2021)通过微博数据分析了新冠疫情背景下公众焦虑情绪的时空分布，表明社交媒体数据能够敏锐地反映社会情绪波动；肖嘉锐等人(2021)通过检索微博数据得到具有代表性的典型文本，从而探讨幼儿入园焦虑与母亲人格特质及焦虑水平。

高考是中国大陆(如无特殊说明，下文中“全国”即指中国大陆部分，共计 31 个省级行政区，不包括未与大陆部分施行同一高考政策的中国港澳台地区)的教考升学制度，同时也是每年必定出现的社会热点事件，而高考改革一直是中国教育领域的热点话题；它不仅关乎教育公平，还涉及广泛的社会心理要素(刘海峰, 2009)。近年来，随着新高考改革的实施，政策的复杂性和区域差异引发了公众的广泛讨论与争议(阎琨、吴菡, 2022; 钟秉林、王新凤, 2019)。例如，张斓(2024)通过分析江苏新高考选科政策的舆情数据，揭示了公众对改革的主要关切点，并提出加强政策沟通与执行优化的建议。

杨现民等(2019)指出，大数据技术在高考改革中发挥了重要作用，包括学生选科决策支持、志愿填报辅助和教学评价优化等方面；大数据技术支持的舆情分析则为政策制定者提供了了解公众需求和意见的有力工具，从而提升了政策的科学性与可行性(张斓, 2024)。此外，钟秉林和王新凤(2019)认为，高考改革的复杂性要求政策制定过程必须以数据为支撑，同时关注公众的心理反应和社会影响。周序(2021)通过“双减”政策下的家长焦虑分析，进一步说明教育政策的实施需要关注社会心理层面的影响，避免引发不必要的公众情绪波动。

情感分析是自然语言处理的重要方向之一，通过分析文本内容可以揭示公众情绪的变化趋势(Wang, 2023)。其中，词典方法是一种基于已有的词汇资源来分

析文本的重要技术，被广泛应用于情感分析、主题建模和语义解析等领域中；利用预先编制的词典，研究者可以借助分词程序实现对海量文本的快速分析。近年来，随着机器学习和自然语言处理技术的不断进步，微博数据的情感分析在准确性和广度上都有了显著提升。然而，数据处理过程中仍然面临诸多挑战，例如数据噪声的过滤、多样化情感表达的识别以及非结构化文本的分析(陈兴蜀等, 2020; 吴泽鹏, 2020)。

常建霞和李君轶(2021)的研究表明，疫情期间微博上的公众焦虑情绪与新增确诊病例数呈显著正相关，进一步证实了情感分析在揭示社会情绪规律中的重要性。张放和甘浩辰(2020)通过分析疫情期间微博情感的时间与空间分布，发现心理距离对公众情绪有显著影响。代一方(2023)强调，准确捕捉微博数据中的情感信息能够为政策调整提供宝贵支持。此外，朱廷劭等(2013)指出，基于社交媒体数据的情感分析还可用于突发事件的风险预警，为社会治理提供了有力工具。

现有的研究在利用社交媒体大数据分析公众情绪与高考政策之间的关系方面仍显不足，尤其是缺乏全国范围和长时间跨度的分析。进行这种大数据分析，需要综合考虑地方高考改革进展、宏观经济指标和应届人数，以及 2020 年后新冠疫情对高考舆情的影响。

本研究结合横向设计与纵向设计，考察新浪微博 2019 年-2024 年中国大陆各省级行政区有关高考的热度变化与文本情感。研究分两阶段进行：研究一将使用 2019-2023 年微博文本数据库、互联网文本爬虫获取 2019-2024 年的 1 月 5 日与 6 月 5 日微博数据集，结合百度指数(index.baidu.com)，考察新高考改革前后可能的热度变化；通过热度数据与宏观指标的统计分析，初步探索影响高考舆情热度的现实因素；在此基础上，研究二将依据大连理工大学情感词典(王大毛, 2021)，使用林萃分析系统进行词切分和情感因素提取，考察高考相关微博文本情感的时空分布，探究其时间变化与疫情、高考改革的关系，以及其空间差异与宏观经济指标的统计关联，尝试解读影响舆论对高考的情感态度的现实因素。

研究一 热度变化

数据获取

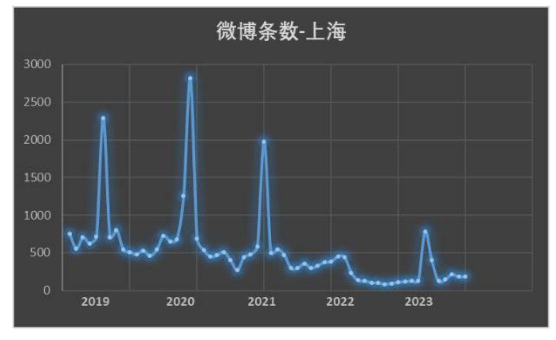
数据集 A：从中国科学院心理研究所社会与工程心理学研究室朱廷劭研究团队构建的 2019-2023 年微博文本数据库中以“高考”为关键词进行提取，删除其中含有“全球高考”(一篇网络小说的标题，在微博平台上具有较高的讨论度)的条目；获得文本数据集 A，共计 425219 条，35661285 字，平均字数为 83.87 字。该数据集来源的微博文本数据库为全时段爬取，具有时空连续性，但由于初始爬取为不完全爬取，并不能覆盖微博全平台所有文本，因此提取出的含“高考”条目量受爬取方法制约，具有一定局限性。

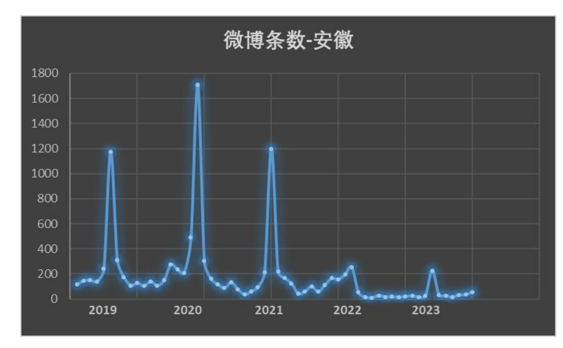
数据集 B：使用 Python 互联网文本爬虫，获取微博全平台 2019-2024 年的 1 月 5 日与 6 月 5 日(高考多为 6 月 7-11 日，选择考前的 6 月 5 日和高考前最后一次寒假前的 1 月 5 日作为爬取对象，共 12 天；下文中“高考季”指每年进行高考和志愿填报的 6-7 月)，含有“高考”、不含“全球高考”的文本；获得文本数据集 B，共计 101842 条，10860655 字，平均字数为 106.64 字。该数据集为单日爬取，不具有时空连续性，反映数据量(热度)的偶然性较大，但实现了对微博全平台单日有关文本数据的全覆盖。

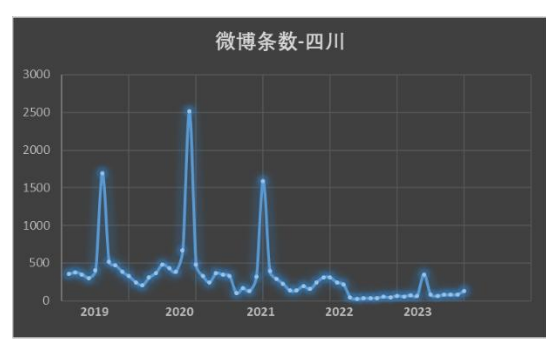
纵向热度趋势

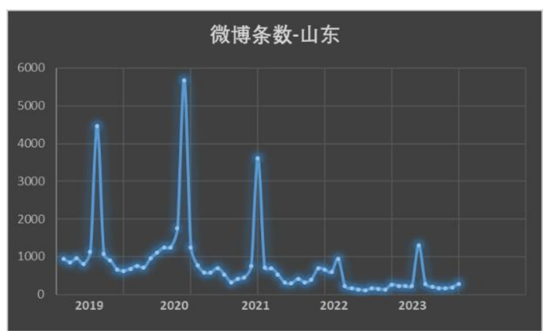
本研究采用文本条目数量作为研究热度的指标。

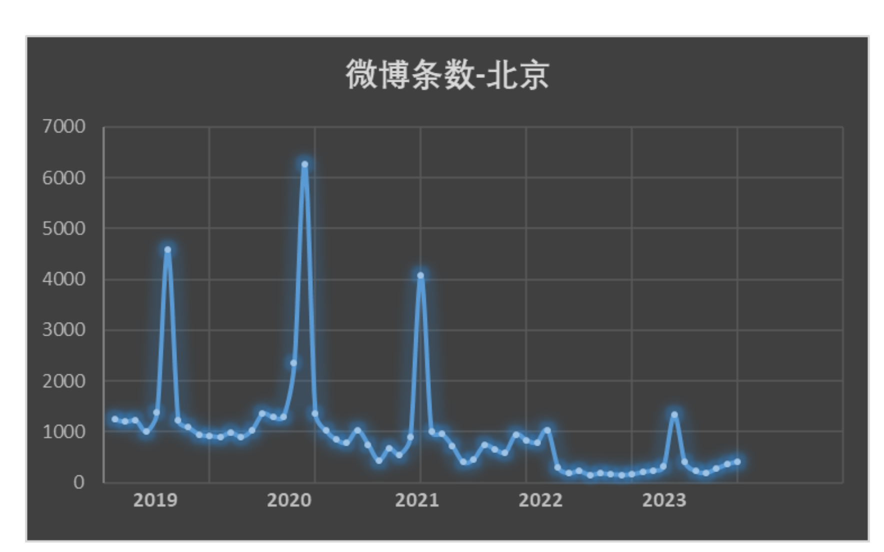
数据集 A 趋势如图：





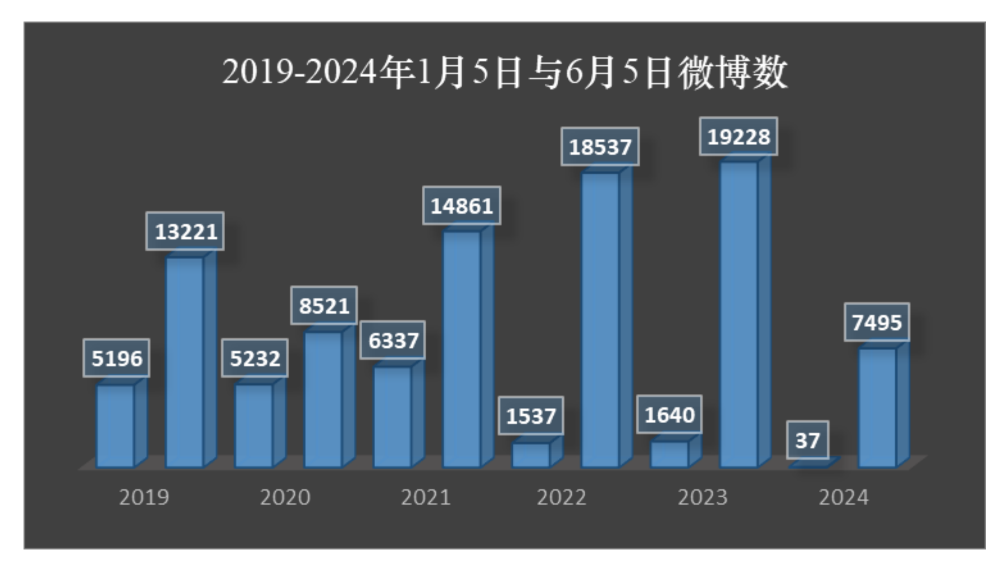






数据集 A 所有省份的条数变化趋势高度同质，且 2022 年、2023 年所有省份均呈现难以解释的低热度，故认为该数据集无法真实反映热度的时空变化；推测可能是由于 A 的原始文本数据库爬取方法所致。

数据集 B 趋势如图：



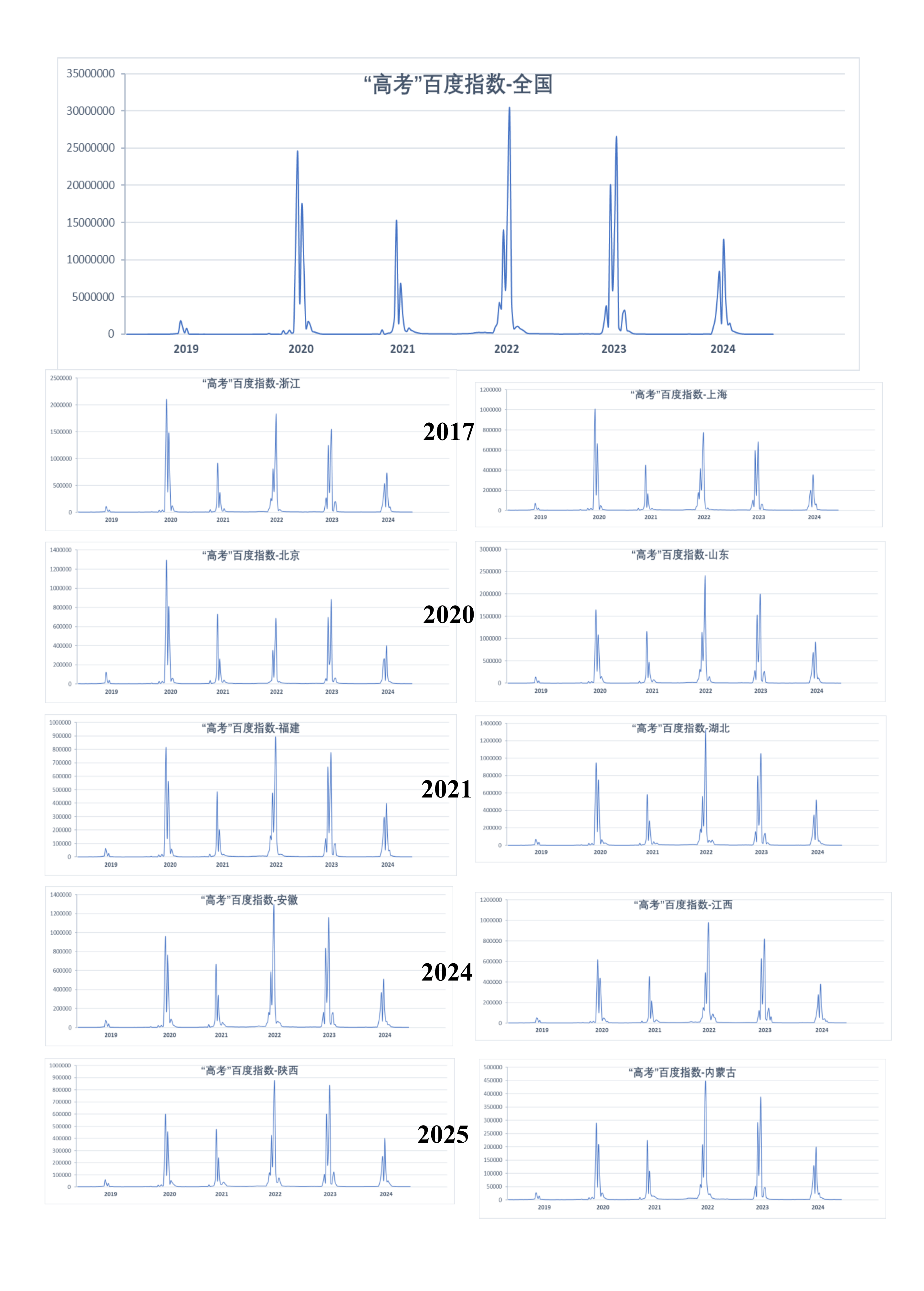
由于单日爬取的局限性，该数据集同样无法真实反映热度的时空变化。

使用**百度指数**(index.baidu.com，基于百度网民行为数据进行综合分析)的**搜索指数**(以网民在百度的搜索量为数据基础，以“高考”关键词为统计对象，分析计算出百度网页搜索中搜索频次的加权和；下文中百度指数均指百度搜索指数)作为反应热度时空变化的指标；

考虑新高考改革施行的时空差异：

批次	实施年份	省份	采用模式
第 1 批	2017 年	上海、浙江	3+3 模式
第 2 批	2020 年	北京、天津、山东、海南	3+3 模式
第 3 批	2021 年	河北、江苏、福建、湖北、湖南、广东、重庆、辽宁	3+1+2 模式
第 4 批	2024 年	吉林、安徽、江西、广西、贵州、甘肃、黑龙江	3+1+2 模式
第 5 批	2025 年	山西、河南、陕西、四川、云南、宁夏、内蒙古、青海	3+1+2 模式

提取每批次两地的百度指数趋势图：



“高考”百度指数-全国
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
2019
2020
2021
2022
2023
2024
“高考”百度指数-浙江
“高考”百度指数-上海
2017
“高考”百度指数-北京
“高考”百度指数-山东
2020
“高考”百度指数-福建
“高考”百度指数-湖北
2021
“高考”百度指数-安徽
“高考”百度指数-江西
2024
“高考”百度指数-陕西
“高考”百度指数-内蒙古
2025

由上图可见，尽管高考改革时间不同，各地搜索指数变化依旧高度同质：2019年高考季指数极低(推测可能由于百度指数算法所需原始数据记录在 2019 年尚不完善)，2022、2023 年较高，2021、2024 年相对较低；同时高考季内部变化呈现明显的双峰，第一峰出现在考试周，第二峰则出现在分数发布、志愿填报的时段；所有省份的搜索指数双峰均呈现 2022 年以前第一峰高于第二峰、2022 年及以后第二峰高于第一峰的现象。

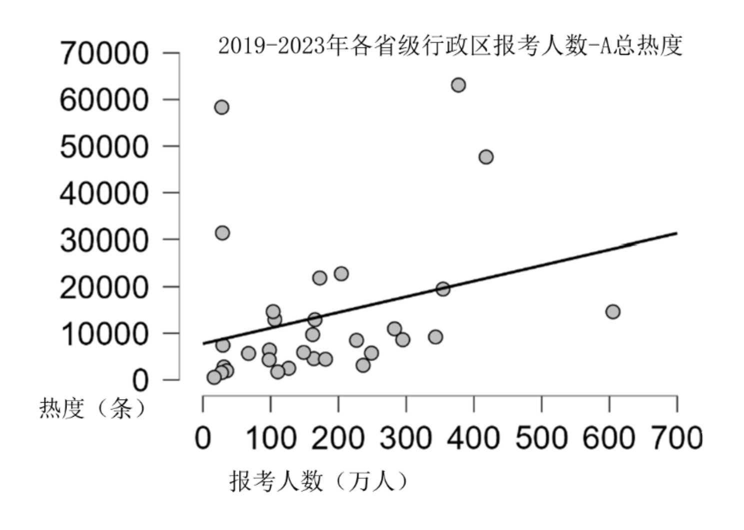
由于百度没有公开搜索指数的具体算法，故难以解释这种同质性的成因；如果算法本身不存在均一化的倾向，那么以上数据可能说明，新高考改革对中文互联网高考热度的影响几乎可以忽略不计；考虑到每年的应届生及其家长群体，其中的大部分可能并不了解高考改革和政策沿袭与变化，这种可能的互联网现象是可以就现实情况进行解释的。换言之，每年高考季热度主要由高考本身的关注度贡献，其互联网数据纵向差异更多地受该时段互联网大环境影响，而非当届与往届的政策差别。

由于六年来各地高考报名人数、财政预算等相关数据均随时间稳定、缓慢增长，难以与随年份无规律变化的热度数据进行相关性检验等统计分析。

横向因素探究

影响高考相关舆情热度空间分布的首要因素应当是地方人数，即各省级行政区每年高考报考人数。通过各地方教育公开网站发布的信息，我们获得了 31 个省级行政区 2019-2024 年高考的报考人数。

取前 5 年总人数数与数据集 A 总热度进行相关性检验，Pearson's r=0.291，



$p=0.112$；观察散点图，注意到回归线之上的三个离群数据，经检查分别为北京、上海、广东，对应“北上广深”四大一线城市，其经济较为发达，信息化程度高；将三地删除后再次进行相关性检验，Pearson's $r=0.547$，$p=0.003$。

取2019-2024年总人数与百度搜索指数，将上述三地删除后进行相关性检验，Pearson's $r=0.834$，$p<0.001$；Spearman's $rho=0.909$，$p<0.001$。 由此可见，百度平台的搜索指数与人数相关性较强，而微博平台的网民构成和文本条目数量受地区发展和互联网使用深度影响较大，具有一定的局限性，可能难以切实反应不同地区高考的现实热度。

显然，经济状况、地方发展会影响教育投入和人民群众的网络使用情况，从而对高考舆情热度产生影响；我们使用2019-2020年的宏观经济指标(总量上，选取地方财政教育支出；人口平均上，选取反映发展水平的人均地区生产总值)与同时段全部31个省级行政区的百度指数总值、数据集A热度总值进行统计学检验：

A热度-地方财政教育总支出：

Pearson's $r=0.718$，$p<0.001$；Spearman's $rho=0.744$，$p<0.001$；

A热度-人均地区生产总值：

Pearson's $r=0.650$，$p<0.001$；Spearman's $rho=0.660$，$p<0.001$；

百度指数-地方财政教育总支出：

Pearson's $r=0.968$，$p<0.001$；Spearman's $rho=0.936$，$p<0.001$；

百度指数-人均地区生产总值：

Pearson's $r=0.302$，$p=0.099$；Spearman's $rho=0.452$，$p=0.011$；

由此可见，同为高考热度总值，相较于来自微博的数据集A，百度指数与地方财政教育支出总量的相关性更强，但与人均地区生产总值相关性较弱；而微博热度则与人均生产总值存在显著的相关性。这也进一步印证了上文的观点：百度平台，特别是其搜索的使用更为普遍，门槛也更低，其搜索指数更多地受网民总量影响；。而微博作为社交媒体，其发布与交互属性决定了其热度受地区发展程度和网络使用深度影响较大，因此与地方财政教育总支出的相关性不如百度指数，但与人均经济数据相关性较强。

研究二 情感成分

数据获取

使用林萃分析系统[1]进行词切分和情感因素提取，获取情感成分的相对值，即每种情感成分在文本内容中的占比[2]。所用词典为大连理工中文情感词汇本体库(王大毛, 2021)；该库是大连理工大学信息检索研究室在林鸿飞教授的指导下经过全体教研室成员的努力整理和标注的一个中文本体资源。该资源从不同角度描述一个中文词汇或者短语，包括词语词性种类、情感类别、情感强度及极性等信息。中文情感词汇本体的情感分类体系是基于 Ekman6 大类情感分类体系；在次基础上，词汇本体加入情感类别“好”对褒义情感进行了更细致的划分，最终词汇本体中的情感共分为 7 大类 21 小类[3]。

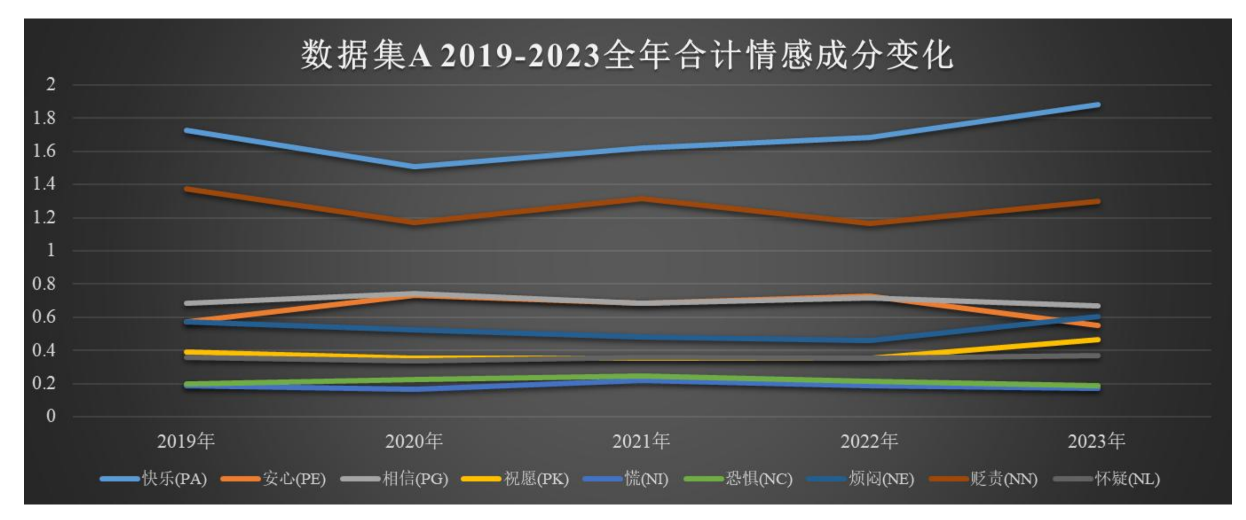
数据集 A 总词数为 35661285；平均字典覆盖率为 0.076；考察 21 小类中全时段全国总值较其余类别较高者，再将分析的结果与无关键词随机抽取的微博文本的情感成分进行对比，选择全国总值较后者同类别差距较大者；最终共有 9 类被纳入考量范围，分别为快乐(PA)、安心(PE)、相信(PG)、祝愿(PK)、慌(NI)、恐惧(NC)、烦闷(NE)、贬责(NN)、怀疑(NL)。

数据集 B 总词数为 10860654；平均字典覆盖率为 0.091；情感成分考察过程及最终获得的考量范围与 A 相同。

时间变化规律

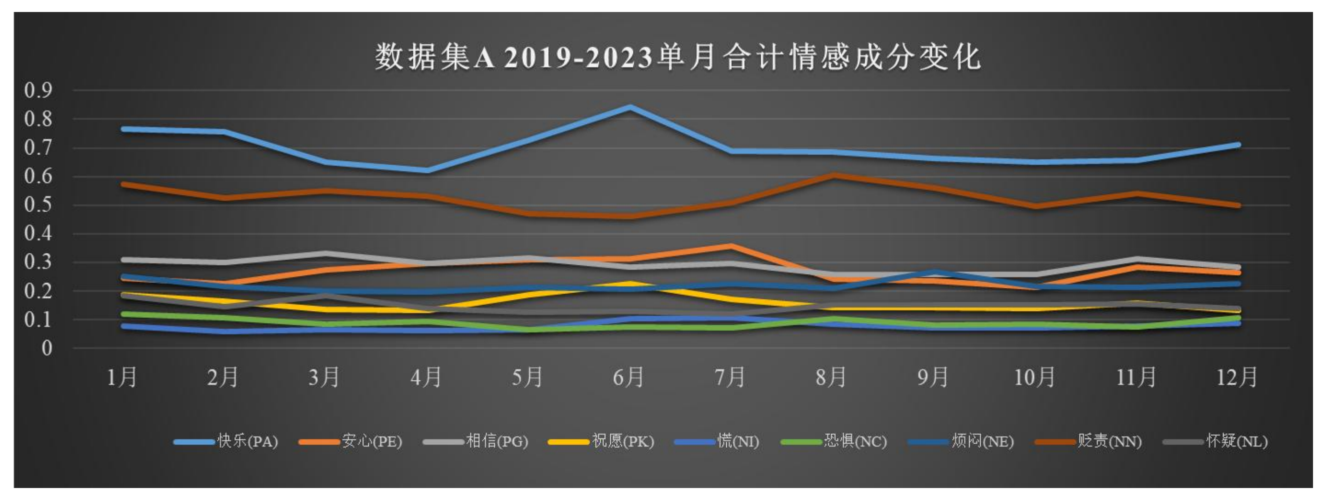
首先考察数据集 A 展示的 2019-2023 全年合计的情感成分变化：

[1] 由中国科学院心理研究所社会与工程心理学研究室朱廷劭研究团队开发并授权使用。
[2] 下文中所用情感成分数据为比例数值的简单求和(无缺失项或成分总值对比)或算数平均(有缺失项或成分间对比)，仅在每张图表内部具有对比意义。
[3] 该本体库及介绍文本取自大连理工中文情感词汇本体库 - Heywhale.com(https://www.heywhale.com/mw/dataset/611383fc911b33001748107d)。



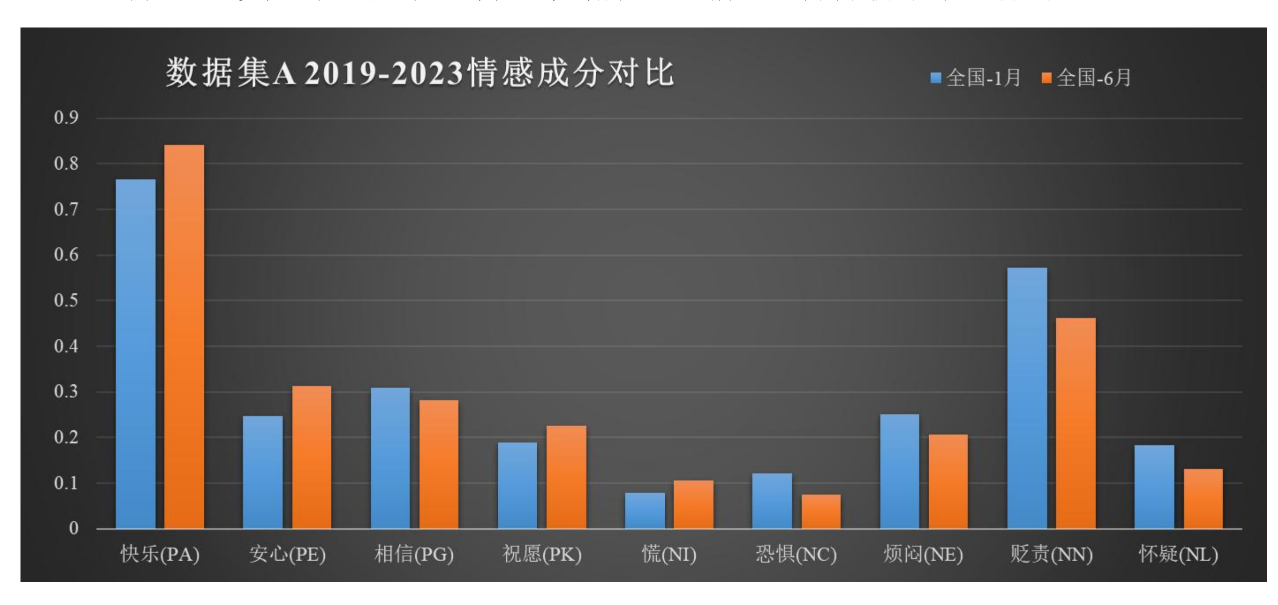
可见快乐(PA)与贬责(NN)具有类似的变化趋势，且与安心(PE)相反。

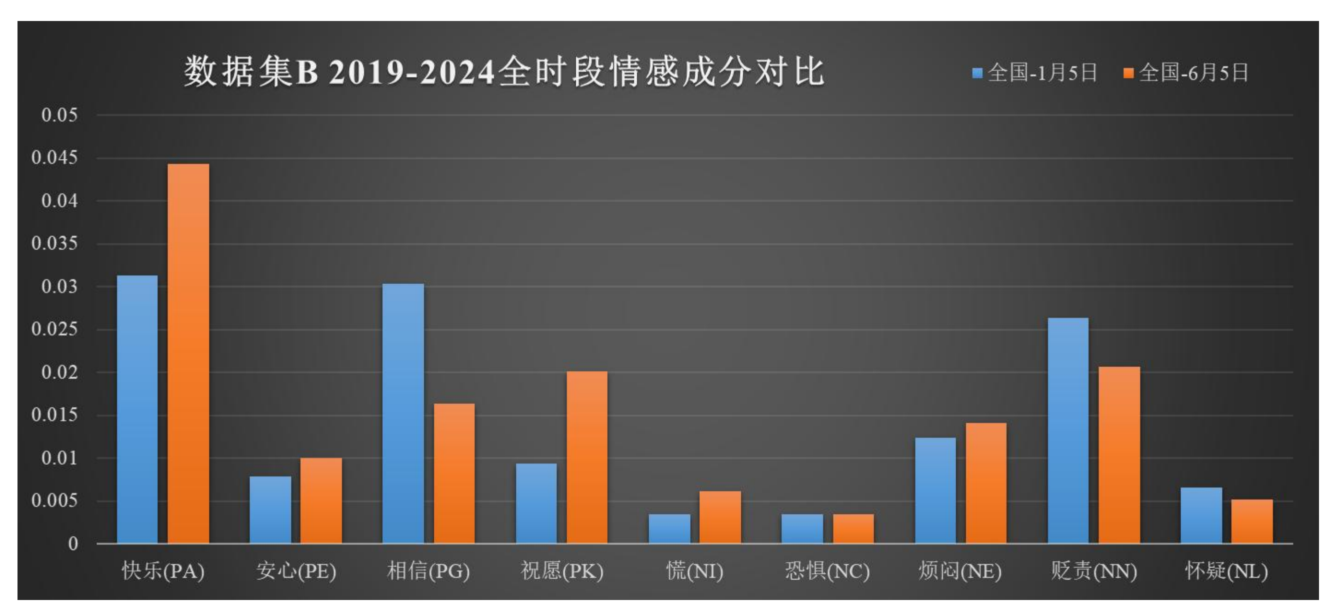
考察数据集 A 展示的 2019-2023 年每年单月合计的情感成分变化：



由此可见，在临近高考的 5 月与进行高考的 6 月，快乐(PA)与祝愿(PK)上升，在考试成绩发布的 8 月贬责(NN)成分较高。

再重点考察 1 月与 6 月，并与数据集 B 的情感成分分析结果进行对比：





由对比图可知，相较于 1 月，马上进行高考的 6 月 5 日、进行和完成高考的 6 月，其文本中快乐(PA)、祝愿(PK)以及慌(NI)的成分均较高，而贬责(NN)和怀疑(NL)的成分较低；这是符合生活常理的。由此可见，微博作为网络媒体和社交平台，文本情感变化依然在某种程度上遵循现实逻辑。

地区差异探究

选取 5 个反映地方经济发展状况与教育投入水平的宏观指标：人均地方生产总值、居民人均可支配收入、人均教育事业费、高中师生比、普通高校师生比，范围为 2019-2020 年，与上述 9 类重点关注的微博文本中的情感成分，按地区切分后进行相关性分析：

数据集 A 2019-2020 年数据：

	人均地区生产总值	全体居民人均可支配收入	人均教育事业费	高中师生比	普通高校生师比
快乐 (PA)					
Pearson's r	0.126	0.146	0.287	-0.103	-0.032
Spearman's rho	0.146	0.175	0.121	-0.059	-0.035
安心 (PE)					
Pearson's r	-0.426*	-0.438*	0.239	0.137	-0.112
Spearman's rho	-0.425*	-0.416*	-0.059	0.172	-0.011
相信 (PG)					
Pearson's r	-0.173	-0.253	0.615***	-0.121	-0.359*
Spearman's rho	-0.17	-0.266	0.265	-0.097	-0.188
祝愿 (PK)					
Pearson's r	0.034	-0.033	0.728***	-0.215	-0.36*
Spearman's rho	0.077	-0.059	0.394*	-0.418*	-0.213
慌 (NI)					
Pearson's r	-0.295	-0.269	0.38*	0.074	-0.376*
Spearman's rho	-0.292	-0.173	0.139	0.076	-0.306
恐惧 (NC)					
Pearson's r	-0.07	-0.135	0.736***	-0.124	-0.136
Spearman's rho	0.029	-0.116	0.517**	-0.089	-0.07
烦闷 (NE)					
Pearson's r	0.078	0.072	0.45*	-0.164	-0.043
Spearman's rho	0.089	0.107	0.11	-0.229	0.089
贬责 (NN)					
Pearson's r	0.296	0.326	0.362*	-0.24	-0.229
Spearman's rho	0.113	0.19	0.345	-0.285	-0.231
怀疑 (NL)					
Pearson's r	0.037	-0.02	0.836***	-0.15	-0.352
Spearman's rho	0.185	0.035	0.653***	-0.282	-0.175

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

由分析结果可见，人均教育事业费是与情感成分关系最密切的；相信(PG)、祝愿(PK)、恐惧(NC)、怀疑(NL)与其呈显著正相关，而慌(NI)、烦闷(NE)、贬责(NN)也有一定相关性。这可能意味着，人均教育事业费越高的地区，其社会与家庭对教育事业越为看重，对于高考的重视程度也相应地高，越可能在网络上表露较为强烈的相信、祝愿、恐惧乃至怀疑等情感，相应地慌张和烦闷的情绪也较为强烈。这一方面通过社会与家庭观念体现在文本中，另一方面则可能是由于教育事业费与经济发展水平乃至网络普及与使用深度存在关联；例如，收入高且重视教育的家庭，其子女教育事业费用高，对高考的重视程度高，且父母与子女在网

络使用深度方面具有经济优势。

此外，安心(PE)与人均生产总值及可支配收入存在负相关，尽管相关性不强，也并不特别显著。这可能是由于经济水平较为落后的地区，其群众对待高考的观念和整体社会态度较为温和；发布的关于高考的社交文本，其背后的想法较为质朴，并不关注各种负面话题，如涉及公平的考试规章和舞弊行为；关注并参与这些舆论可能引起相信、恐惧、怀疑等情感，从而降低安心的成分占比。

数据集 B 由于 2019-2020 年的四天数据过少，使用全部数据，合并后进行相关性分析：

	人均地区生产总值	全体居民人均可支配收入	人均教育事业费	高中师生比	普通高校生师比
快乐 (PA)					
Pearson's r	-0.18	-0.137	0.136	-0.067	0.054
Spearman's rho	-0.153	-0.138	0.151	-0.082	0.037
安心 (PE)					
Pearson's r	0.682***	0.711***	0.256	-0.472*	-0.514**
Spearman's rho	0.577**	0.477*	0.211	-0.325	-0.396
相信 (PG)					
Pearson's r	0.353	0.397*	0.242	-0.117	-0.046
Spearman's rho	0.406*	0.487**	0.104	-0.122	-0.175
祝愿 (PK)					
Pearson's r	0.104	0.128	0.031	0.028	-0.065
Spearman's rho	0.266	0.321	0.156	-0.063	-0.176
慌 (NI)					
Pearson's r	0.242	0.206	0.095	-0.024	-0.161
Spearman's rho	0.473*	0.455*	0.162	-0.116	-0.272
恐惧 (NC)					
Pearson's r	-0.132	-0.122	-0.123	0.419*	0.032
Spearman's rho	0.101	0.068	0.046	0.091	-0.066
烦闷 (NE)					
Pearson's r	0.176	0.083	-0.119	0.088	-0.016
Spearman's rho	0.452*	0.408*	-0.047	0.18	0.003
贬责 (NN)					
Pearson's r	-0.047	-0.058	-0.319	0.199	0.045
Spearman's rho	0.033	0.096	-0.36*	0.102	0.049
怀疑 (NL)					
Pearson's r	-0.22	-0.269	0.306	0.081	-0.165
Spearman's rho	-0.185	-0.287	0.003	0.091	-0.295

*p < 0.05, **p < 0.01, ***p < 0.001

由上述分析可见，安心(PE)与前两项存在显著正相关，而人均教育事业费与

几乎所有情感成分均无显著相关性。如此看来，A 中的相关性结果似乎并不稳固，但由于 B 为单日爬取，偶然因素多，随机性较大；且 B 是 2019-2024 年的平均数据，与 2019-2020 年宏观数据进行相关性分析，带有一定的预测性质，故上述分析结果难以推翻 A 分析所得的结论。

讨论

本研究综合使用社交媒体数据、情感分析和地方宏观指标，系统探讨了高考舆情热度及情感成分的时空变化规律，部分弥补了以往研究中时间跨度不足、空间分布不全的不足。通过对微博大数据的文本处理与情感分析，我们发现，不同平台(百度、微博)、不同地区的用户认知倾向和互联网使用习惯不同，影响热度和情感表达；而地区经济发展水平、教育投入和网络使用深度等因素也影响着高考相关舆情的空间分布和情感成分。本研究所采用的多数据源、多样本数据集分析对照方法，也为未来研究社会重大公共议题的舆情规律提供了借鉴。

具体而言，对于热度的纵向研究，数据集 A 和 B 受爬取方法限制，无法真实反映热度的时间变化。而百度指数作为反映热度时间变化的可靠指标，其各地搜索指数变化却高度同质；这可能说明，高考在百度平台的热度主要由高考本身的关注度驱动，受新高考改革影响较小。这也提醒我们，新高考改革的推进需要充分考虑地方经济水平与社会情绪，在宣传及舆论工作方面有的放矢，以确保政策实施的公平性和公众的接受度。热度的横向研究则显示，报考人数、地方财政教育支出与百度指数相关性较强，与数据集 A 相关性较弱；人均地区生产总值则对两个数据来源相反。这可能反映了不同平台的用户群体差异。对于微博文本数据的情感成分分析表明，情感成分的变化趋势符合高考这一特定社会热点的现实逻辑，体现了公众情绪在重大事件中的波动。地区经济和教育投入，但不同情感类别的影响模式存在差异。例如，“人均教育事业费”这一指标显著影响高考相关舆情的情感成分，而经济水平较高地区安心(PE)成分较低，可能与其认知特点有关。

尽管研究揭示了舆情热度及情感成分的诸多规律，但仍存在若干问题。首先，微博文本的采集方法与分析策略可能导致部分数据偏差，百度指数算法不透明、

爬虫功能可能存在的问题以及现有数据库的爬取思路等均限制了研究微博文本数据的“无偏取样”；尤其是数据集 A 中 2022 年与 2023 年的条目异常，使得通过 A 得到的情感分析结果的可靠性和可解释程度从侧面遭到削弱，其与宏观指标的纵向分析也受到极大限制。

此外，情感分析依赖词典覆盖率，本研究所用文本的词典覆盖率总体上在 0.1 以下，部分时空范围文本的词典覆盖率接近 0.05，其分析的可靠性不足。另一方面，在网络语言变幻无常、日新月异，青少年群体“内部语言”层出不穷的今天，基于现实生活编制的情感词库对微博文本和高考特定语境的适用性可能有限。

未来针对高考等周期性社会热点议题的大数据研究，应当进一步扩展数据来源，例如结合微信、小红书、知乎等互联网社交平台的数据，探索不同用户群体对于高考的热度与情感成分差异；可以使用机器学习模型对情感成分进行更精细的预测，以提高情感分析的准确性与广度。同时，如果可以进行多平台跟踪研究，探讨更长时间跨度、更多社交媒体的高考舆情与地方发展、社会情绪的关联，构建更加完整的研究框架，就能进一步理解高考舆情的形成机制及其与社会经济因素的关系，并为相关政策的制定提供更有力的数据支持。

生成式人工智能技术应用声明：在本文撰写结束后，作者使用 ChatGPT-4o 对部分文本进行了润色，已根据需要对生成的内容进行了审查和编辑，并手动修改至论文中。

参考文献

Wang, R. (2023). Sentiment Analysis Technology of Chinese Weibo under the Background of Big Data.

Xiao, Y., Li, B.,、 Gong, Z. (2018). Real-time identification of urban rainstorm waterlogging disasters based on Weibo big data. *Natural Hazards*, *94*(2), 833-842.

Zhao, X.,、Wang, X. (2023). Dynamics of Networked Framing: Automated Frame Analysis of Government Media and the Public on Weibo With Pandemic Big

Data. *Journal Mass Commun Q*, *100*(1), 100-122.

常建霞,、李君轶. (2021). 新冠肺炎疫情和公众焦虑情绪的时空分异研究——基于微博数据的分析. *人文地理*, *36*(03), 47-57+166.

陈兴蜀, 常天祐, 王海舟, 赵志龙,、张杰. (2020). 基于微博数据的"新冠肺炎疫情"舆情演化时空分析. *四川大学学报(自然科学版)*, *57*(02), 409-416.

代一方. (2023). 基于微博数据的人工智能网络舆情分析——以 ChatGPT 话题为例. *传播与版权*(21), 93-95.

杜治娟, 王硕, 王秋月,、孟小峰. (2017). 社会媒体大数据分析研究综述. *计算机科学与探索*, *11*(01), 1-23.

范晓磊. (2023). *突发公共事件下基于微博数据的舆情分析* [硕士]

黄峰, 李赫, 丁慧敏, 吴胜涛, 刘明明, 刘天俐, 刘晓倩,、朱廷劭. (2020). 社会媒体大数据视角下经济发展对集体道德的影响模式. *科学通报*, *65*(19), 2062-2070.

刘海峰. (2009). 高考改革的思路、原则与政策建议. *教育研究*, *30*(07), 3-7.

王大毛. (2021). *大连理工中文情感词汇本体库*

吴胜涛, 茅云云, 吴舒涵, 冯健仁, 张庆鹏, 谢天, 陈浩,、朱廷劭. (2023). 基于大数据的文化心理分析. *心理科学进展*, *31*(03), 317-329.

吴泽鹏. (2020). 大数据分析的困境及语境论视域下的解决思路. *学理论*(03), 60-61.

肖嘉锐, 王康慧, 刘明明, 上官芳芳,、朱廷劭. (2021). 幼儿入园焦虑与母亲情绪特征的关系:基于新浪微博的研究. *中国妇幼卫生杂志*, *12*(03), 11-15.

阎琨,、吴菡. (2022). 新高考改革的现实困境与突破策略. *中国高教研究*(02), 13-20.

杨现民, 郭利明, 晋欣泉,、顾佳妮. (2019). 大数据助力新高考改革:框架设计与实施路径. *电化教育研究*, *40*(02), 30-37.

张放,、甘浩辰. (2020). 疫情心理时空距离对公众情绪的影响研究——基于新冠肺炎疫期微博文本面板数据的计算分析. *新闻界*(06), 39-49.

张斓. (2024). 网络舆论与高考政策互动关系研究——基于江苏新高考选科舆情处置案例分析. *中国考试*(05), 28-35.

钟秉林,、王新凤. (2019). 新高考的现实困境、理性遵循与策略选择. *教育学报*, *15*(05), 62-69.

周序. (2021). 家庭资本与学业焦虑——试论“双减”政策引发的家长焦虑问题. *广西师范大学学报(哲学社会科学版)*, *57*(06), 96-106.

朱廷劭, 白朔天,、李昂. (2013). 基于社会媒体大数据的群体事件风险预警. 心理学与创新能力提升——第十六届全国心理学学术会议, 中国江苏南京.

(通讯作者: 朱廷劭 Email:tszhu@psych.ac.cn)

Public Opinion Research on College Entrance Examination Based on Weibo Big Data: Temporal and Spatial Changes of Heat and Emotional Components and Analysis of the Influence of Macro Factors

Dong Jincheng[1, 2] Ai Yuheng[3, 4] Zhu Tingshao[1, 2]

[1] (Institute of Psychology, Chinese Academy of Sciences, Beijing 100101)

[2] (Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049)

[3] (Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

[4] (School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049)

Abstract Based on the textual content of Weibo big data, this study analyzes the changes in the heat and sentiment components of public opinion about the college entrance examination in various provincial-level administrative regions in mainland China between 2019 and 2024, and explores the relationship between these changes and macroeconomic indicators. The study used the number of text entries as the heat index and analyzed it with Baidu Index, and the results showed that the heat of public opinion related to the college entrance examination varied across years and provinces, and was closely related to local economic indicators such as per capita gross regional

product and local financial expenditure on education. And the sentiment component analysis with the help of the Linxian analysis system and the Dalian Polytechnic Sentiment Dictionary shows that the texts related to the college entrance examination show different component characteristics in different time periods and regions, and are associated with some realistic factors. This study provides a new big data perspective for understanding the changing patterns of the heat and sentiment components of public opinion on the college entrance examination, and its relationship with socioeconomic factors.
Keywords Big data; social media; microblogging; college entrance examination; heat data analysis; sentiment component analysis